Description phonologique et codification économique du langage

Gabriel G. Bès

Description phonologique et codification économique du langage

Gabriel G. Bès
Groupe de recherches sur la condensation de l'information en langue naturelle (CILN)
Université Blaise-Pascal, Clermont II

Cahiers du Centre Interdisciplinaire des Sciences du Langage, 1980, n° 2, p. 117-123

Résumé

Pour transmettre et/ou stocker un texte en langue naturelle, chaque symbole de son signifiant est codifié par une suite de symboles binaires. Dans le cadre de la codification économique du langage, on se propose de réduire autant que possible le nombre de symboles binaires utilisés tout en sauvegardant l'univocité de la codification et de la décodification. Ce texte présente *le Code de réduction alphabétique C*, où les symboles sont organisés en une pluralité de systèmes en fonction de leurs contextes d'apparition. Un tel code témoigne à la fois d'une manière d'explorer les caractéristiques structurelles des langues naturelles productrices de redondance et de la possibilité d'apporter une réponse, à partir des caractères propres aux langues naturelles, au problème de la condensation de l'information.

Voir aussi

Gabriel G. Bès. *Identités et différences dans les unités de deuxième articulation.* Thèse. Université René Descartes - Paris V, 1972. https://hal.archives-ouvertes.fr/tel-01095229

Gabriel G. Bès. « Structuration linguistique et mesure de l'information. » *Linguistique fonctionnelle. Débats et perspectives*, présentés par M. Mahmoudian. Presses universitaires de France, Paris, 1979, p. 129-141. https://hal.archives-ouvertes.fr/hal-01100168

Gabriel G. Bès. « Un théorème sur l'équivalence de la valeur H entre langages ». *Condenser*, Adosa, Clermont-Ferrand, février 1980, n° 1, p. 97-100. https://hal.archives-ouvertes.fr/hal-01100224

Cahiers du C.I.S.L, n°2, 1980,
Université Toulouse le Mirail.

Description phonologique et codification économique du langage

par Gabriel G. Bès *

On sait que, pour transmettre et/ou stocker un texte en langue naturelle, chaque symbole de son signifiant est codifié par une suite de symboles binaires. Dans le cadre de la codification économique du langage, on se propose de réduire autant que possible le nombre de symboles binaires utilisés tout en sauvegardant l'univocité de la codification et de la décodification.

Cette question a été abordée dans le cadre de la théorie de l'information : pour mesurer la valeur d'entropie d'un système de symboles, on procède par approximations successives. Dans une approximation d'ordre 0, on ne tient compte que du nombre de symboles à distinguer ; dans une approximation d'ordre 1, on considère la fréquence des symboles isolés ; dans une approximation d'ordre 2, la fréquence des symboles dans le cadre du digramme ; dans une approximation d'ordre 3, la fréquence des symboles dans le cadre du trigramme, et ainsi de suite. A chaque nouvelle approximation, on obtient une valeur d'entropie qui est égale ou inférieure à celle de l'approximation précédente. Le nombre de symboles binaires utilisés pour codifier un système de symboles est égal ou supérieur à la valeur d'entropie du système.

* Professeur à l'Université de Clermont II

En considérant des blocs de lettres de plus en plus grands, on peut donc diminuer la valeur d'entropie et, par là, réduire le nombre de symboles binaires nécessaires à la codification, mais, par là même, on augmente la complexité du système de codification. Dans le cas d'un texte écrit en langue naturelle, si l'on considère que l'alphabet est constitué de 26 lettres plus un espace (27 symboles) pour calculer la valeur d'entropie dans une approximation d'ordre 2, il faut disposer de 27^2 valeurs (la fréquence de chaque symbole à la suite de chacun des symboles de l'alphabet), de 27^3 valeurs, dans une approximation d'ordre 3 (la fréquence de chaque symbole à la suite de chacun des 27^2 diagrammes possibles). En pratique, la complexité du système de codification devient ainsi vite insurmontable.

Pour aborder cette question à partir d'une optique linguistique, on a considéré que les phonèmes d'une langue ne constituent pas un alphabet unique mais que, au contraire, ils s'organisent dans une pluralité de systèmes. Le problème central est donc de codifier les symboles appartenant à chaque système. Par exemple, sur le plan graphique, dans une langue comme l'espagnol, les p des mots pata, apto et capa appartiennent à des systèmes différents, et doivent, par conséquent, être codifiés différemment.

Dans les langues naturelles, les contextes qui conditionnent les systèmes sont souvent constitués par des éléments qui apparaissent, dans la chaîne du signifiant, après les éléments du système considéré. Toujours en espagnol, le s de la suite

cas...
 y

appartiendra au système des consonnes intervocaliques si, dans la position y, on trouve une voyelle ; il appartiendra au système des consonnes de fin de syllabe si, dans y, on trouve une consonne.

Pour assurer l'univocité de la décodification, les éléments indicateurs du contexte doivent être organisés linéairement, de gauche à droite,de manière qu'à partir de chaque symbole de la chaîne on sache à quel système appartient le symbole suivant. Dans l'exemple proposé, à partir du a de cas..., on doit posséder l'information sur le caractère intervocalique ou de fin de syllabe du s qui suit. Cette organisation de l'information est obtenue moyennant un ensemble de règles permettant, à partir d'un texte en langue naturelle, d'obtenir un"langage organisé", où l'information est présentée de gauche à droite. On exige que le texte en langue naturelle et le langage organisé qui, par règle, lui est associé, soient en relation bijective.

Dans ce cadre d'ensemble, caractérisé par la pluralité des systèmes considérés et par l'organisation linéaire de l'information dans le langage organisé qui est associé à chaque texte, on a obtenu quelques résultats intéressants. Le Code de réduction alphabétique C, appliqué à l'espagnol, est constitué de 273 symboles au total (qui correspondent aux lettres et à l'espace dans les différents contextes) distribués dans 14 systèmes différents. Si l'on accepte que la valeur en symboles binaires nécessaires pour codifier un texte est égale à la valeur d'entropie du système, il est possible de codifier l'espagnol au moyen de 3,03 à 3,08 symboles binaires par lettre et espace. Ce résultat est à comparer à la valeur de 3,1 obtenue par Shannon pour l'anglais dans une approximation d'ordre 3, qui, rappellons-le, nécessite que l'on dispose de 27^3 valeurs. Le Code de réduction alphabétique C a été appliqué manuellement à un corpus réduit de textes en espagnol (env. 3.000 lettres et espaces); ses résultats doivent, par conséquent, être vérifiés sur un corpus plus large, même s'ils semblent assez solidement établis. D'ailleurs, ces résultats ne représentent certainement pas l'économie optimale qu'on pourrait obtenir dans une approche de ce type.

Sur un plan général, il a été possible de démontrer que les langages organisés ne possèdent pas nécessairement une valeur d'entropie plus élevée que celle des textes en langue naturelle auxquels ils sont associés, même si l'alphabet du langage organisé comporte plus de symboles que celui du texte en langue naturelle auquel il est associé. Plus précisement, on a montré que

$$H \cdot N(1) = H' \cdot N(1')$$

où

H = entropie optimale de L

$H \triangleq H_i$ de L

H'= entropie optimale de L'

$H' \triangleq H_i$ de L'

$N(1)$ = nombre donné de lettres et/ou espaces dans une suite finie de L.

$N(1')$= nombre donné de lettres et/ou espaces dans une suite finie de L'.

Les codes de réduction alphabétique ainsi envisagés témoignent à la fois d'une manière d'explorer les caractéristiques structurelles des langues naturelles productrices de redondance et de la possibilité d'apporter une réponse, à partir des caractères propres aux langues naturelles, au problème de la condensation de l'information.

Note

Le texte précédent est un résumé de la communication présentée aux Journées de linguistique organisées par le Centre Interdisciplinaire des Sciences du Langage de l'Université de Toulouse-Le Mirail (mars-avril 1979); le matériel complémentaire peut être trouvé dans G.G.Bès "Structuration linguistique et mesure de l'information", dans M. Mahmoudian [Ed.] Linguistique fonctionnelle, débats et perspectives.Paris, P.U.F., 1979, p. 129-141; G.G. Bès "Un théorème sur l'équivalence de la valeur de H entre langages", dans

Condenser, 1 (février 1980), p. 97-100; G.G.Bès et D.Guillot Code de réduction alphabétique A. Mendoza, UNC-UTN, 1974. La constitution des systèmes partiels d'un système phonologique a été discutée dans G.G.Bès Identités et différences dans les unités de deuxième articulation (Thèse, Paris, 1972). On trouvera ci-joint le tableau récapitulatif des résultats obtenus pour les 14 systèmes du langage organisé dans le Code de réduction alphanétique C. Col. I (Occur.): nombre total d'occurrences des symboles de chaque système; Col. III ($f(S_i)$): fréquence de chaque système; V ($H(S_i)$): entropie de chaque système; Col.VII ($f(S_i) \cdot H(S_i)$): fréquence de chaque système multipliée par son entropie.

Le corpus étant réduit, certaines lettres de certains systèmes n'ont pas été attestées, même si elles sont qualitativement possibles. Pour pallier cet inconvénient, les résultats obtenus ont été corrigés de la manière suivante:

Y = nombre total d'occurrences attestées dans un système donné

$$Y = \frac{99}{100} \cdot Y'$$

Y'= nombre total d'occurrences corrigées

$\frac{1}{100} \cdot Y'$ a été distribué entre les symboles non attestés de telle manière que, dans chaque système, la fréquence corrigée de chacun des symboles non attestés est égale à:

$$\frac{0{,}01 \cdot Y'}{\text{Nombre de symboles non attestés}}$$

Les résultats qu'on obtient en tenant compte des occurrences corrigées apparaissent dans les colonnes II, IV, VI et VIII.

En bas du tableau sont indiqués le nombre total d'occurrences (Col.I), le nombre total d'occurrences corrigées (Col.II), la valeur moyenne, en bits, de chaque symbole, sans correction (Col. VII) et avec correction (Col. VIII).

Les valeurs indiquées dans le tableau sont celles des symboles du langage organisé. Comme un symbole de ce langage peut correspondre, dans certains cas, à une suite d'un ou plusieurs symboles (lettres et/ou espaces) du

texte en langue naturelle auquel il est associé, le nombre de symboles dans le langage organisé est plus réduit que le nombre des lettres et espaces du texte original:

Occurrences dans le texte original : 2.940

Occurrences dans le langage organisé: 2.676

On obtient par conséquent les valeurs suivantes de H pour le corpus original:

Sans correction : $\frac{2.676}{2.940} \cdot 3,326 = 3,027$ Bits

Avec correction : $\frac{2.676}{2.940} \cdot 3,378 = 3,073$ Bits

S_i	Occur. I	Occ. cor. II	$f(S_i)$ III	$f(S'_i)$ IV	$H(S_i)$ V	$H(S'_i)$ VI	$f(S_i).H(S_i)$ VII	$f(S'_i).H(S'_i)$ VIII
1	508	513,08	0,190	0, 190	4, 160	4, 231	0,790	0,804
2	46I	465,61	0, 172	0, 172	3, 235	3,235	0,556	0,556
3	63	63,63	0,023	0,023	2,207	2,285	0,051	0,052
4	32	32,32	0,012	0,012	2,162	2,237	0,026	0,027
5	416	420,16	0,155	0,155	3,482	3,482	0,540	0,540
6	27	27,27	0,010	0,010	3,38	3,437	0,034	0,034
7	490	494,9	0,183	0,183	2,651	2,742	0,485	0,502
8	128	129,01	0,048	0,048	2,860	2,952	0,137	0,142
9	25	25,25	0,009	0,009	1,225	1,304	0,011	0,012
10	392	315,12	0,117	0,117	4,108	4,163	0,481	0,487
11	100	101	0,037	0,037	3,456	3,533	0,128	0,131
12	33	33,33	0,012	0,012	1,860	1,961	0,022	0,023
13	56	56,56	0,021	0,021	1,946	2,043	0,041	0,043
14	25	25,25	0,009	0,009	2,697	2,784	0,024	0,025

T.oc. : 2,676

T.oc.cor.: 2702,49

$-\sum_{i=1}^{i=n} f(S_i).H(S_i) = 3,326$

$-\sum_{i=1}^{i=n} f(S'_i).H(S'_i) = 3,378$